AAAI Technical Report FS-12-06
Machine Aggregation of Human Judgment

Improving Predictions with Hybrid Markets

Yiftach Nagar and Thomas W. Malone
MIT Center for Collective Intelligence and MIT Sloan School of Management
100 Main St., Cambridge, MA, USA
{nagar, malone} @mit.edu

Abstract

Statistical models almost always yield predictions that are more accurate than those of human experts. However, humans are better at data acquisition and at recognizing atypical circumstances. We use prediction markets to combine predictions from groups of humans and artificial-intelligence agents and show that they are more robust than those from groups of humans or agents alone.

Introduction

How can we make predictions about actions or behaviors in complex social systems? Recent advances in artificial-intelligence enable artificial agents to relatively successfully identify patterns even in complex scenarios (e.g. Jain, Duin, & Mao, 2000; Mannes et al., 2008). Substantial evidence from multiple domains suggests that models usually yield better (and almost never worse) predictions than do individual human experts (Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Whereas models (or machines) are better at information processing and are consistent (Einhorn, 1972), humans suffer cognitive and other biases that make them bad judges of probabilities (Kahneman & Tversky, 1973; Lichtenstein, Baruch, & Phillips, 1982). In addition, factors such as fatigue can produce random fluctuations in judgment (Dawes, et al., 1989). Indeed models of judges often outperform the judges themselves (Armstrong, 2001b; Goldberg, 1970; Stewart, 2001). When working in groups, humans often exhibit effects such as groupthink (Janis, 1972) and group polarization (Brown, 1986) that negatively affect their judgment. Nevertheless, humans still have an important role in predicting real-life situations,, for at least two good reasons. First, humans are still better at tasks requiring the handling of various types of information – especially unstructured information – including retrieval and acquisition (Einhorn, 1972; Kleinmuntz, 1990), categorizing (von Ahn & Dabbish, 2004), and pattern recognition (Mitra et al., 2009; von Ahn, Blum, Hopper, & Langford, 2003). Second, humans' common-sense is required to identify and respond to "broken-leg" situations (Meehl, 1954) in which the rules normally characterizing the phenomenon of interest do not hold. Therefore, combining human and machine predictions may help in overcoming the respective flaws of each. The scarcity of both theoretical and empirical work to that end is conspicuous. Previous work (Blattberg & Hoch, 1990; Bunn & Wright, 1991; Einhorn, 1972) emphasized the complementary nature of humans and models, but did not stress the potential of improving predictions by combining predictions from multiple humans and models. We know, however, that combining forecasts from multiple independent, uncorrelated forecasters leads to increased forecast accuracy whether the forecasts are judgmental or statistical (Armstrong, 2001a; Makridakis, 1989; Winkler, 1989). Further, because it may be difficult or impossible to identify a single forecasting method that is the best, *"it is less risky in practice to combine forecasts than to select an individual forecasting method"* (Hibon & Evgeniou, 2005). We conjecture, therefore, that in situations where rules are fuzzy or difficult to discern, and where some data are hard to codify, combining predictions from groups of humans and artificial-intelligence agents together can yield predictions that are more accurate and robust than those created by groups of either type alone.

This paper is a shorter version of our original paper, which appeared in the proceedings of The Thirty Second International Conference on Information Systems (ICIS 2011), Shanghai, China.

The Potential Advantages of Markets

But how to combine? Many different ways of combining predictions are explored in the literatures of forecasting and model fusion, including simple and weighted averaging, majority voting, max, min, median, etc., as well as techniques that involve learning, e.g. Bayesian learning. Theoretical and empirical comparisons have shown that no single method or rule of combination are best under all circumstances (see, for example, Armstrong, 1989; Clemen, 1989; Duin & Tax, 2000; Kuncheva, 2002; Lam, 2000). The simple average is usually considered a good default (Armstrong, 2001a), and the most robust against 'classifier peculiarities' (Lam, 2000).

Over the past decade, following the success of prediction markets[1] in public settings (Berg, Forsythe, Nelson, & Rietz, 2001), many companies have started using them to efficiently aggregate predictions from employees (Cowgill, Wolfers, & Zitzewitz, 2008; Malone, 2004; Sunstein, 2006; Surowiecki, 2004; Wolfers & Zitzewitz, 2004). Empirical investigations of prediction markets performance have shown that indeed they yield predictions that are usually at least as accurate and as calibrated as other methods traditionally used for forecasting (Chen, K.-Y. & Plott, 2002; Cowgill, et al., 2008; Hopman, 2007; Ortner, 1997; Spann & Skiera, 2009). Prediction markets also fared well against other methods of combining predictions such as simple average, weighted average and a logarithmic regression (Berg, Nelson, & Rietz, 2008; Chen, Y., Chu, Mullen, & Pennock, 2005).

While prediction markets have mostly been used to aggregate predictions from humans, there is no reason why the mechanism cannot be used to aggregate predictions from software agents; yet this option remains mostly unexplored[2]. One exception is a recent study by Perols et al. (2009) who used a prediction market to combine predictions from machine classifiers. In their experiment, depending on the setting, the market mechanism either outperformed or performed on par with 3 benchmark combination mechanisms: simple average, weighted average, and majority. To the best of our knowledge, no one, to this day, has tried to use prediction markets for combining human and model predictions.

It is certainly possible that, in some scenarios, prediction markets will provide only minute improvements in accuracy over other methods, as two recent studies (Goel, Reeves, Watts, & Pennock, 2010; Perols, et al., 2009) suggest; and costs of implementation, set-up and training should be considered as well. However, prediction markets may be appealing in some settings for reasons beyond accuracy improvement. First, as Perols et al. (2009) note, unlike some combination methods that require learning and setup, prediction markets can adjust to changes in base-classifier composition and performance without requiring offline training data or a static ensemble composition reconfiguration. Second, by increasing attentive participation and by tying compensation to performance, while giving participants a sense of both fun and challenge, they serve to increase both extrinsic and intrinsic motivation. For instance, human participants in prediction markets have an economic incentive to gather more information that would improve their performance in the markets. Third, the use of markets can also help knowledge discovery and sharing in organizations (Cowgill, et al., 2008; Hayek, 1945), especially so in large, distributed and/or virtual environments. Fourth, they also induce a sense of participation which supports the legitimacy and acceptance of the predictions made. Finally, Markets can also be open for people to design and run their own 'pet' agents, thus potentially incorporating an open, continuous improvement pattern into the forecasting process. For these reasons, prediction markets are a potentially useful and appealing mechanism for dynamically combining predictions from a varying population of humans and agents in real organizational settings.

We hypothesize, therefore, that combining predictions made by humans and artificial-intelligence agents can outperform both predictions made solely by humans or solely by artificial-intelligence (or statistical) models. We also hypothesize that prediction markets can be a useful mechanism for dynamically combining human and agent predictions.

It is important to realize that we are not claiming that combining human and machine predictions in this way is *always* better, only that it is *sometimes* better. As such, our results can be seen as an existence proof of one situation in which it is better. In the conclusion of the paper below, we speculate about the range of other situations in which this approach may be superior.

Method

To test these hypotheses, we conducted a study whose goal was to compare the quality of predictions made by three different types of predictors: groups of humans, groups of artificial-neural-network agents, and 'hybrid' groups of humans and agents. We used prediction markets to combine the predictions that these three types of groups made. In each case, the groups predicted whether the next play in an American football game would be a "run" or

[1] Also known as information markets, decision markets, electronic markets, virtual markets, idea futures, event futures and idea markets (Tziralis & Tatsiopoulos, 2007; Wolfers & Zitzewitz, 2004)

[2] Albeit some researchers in machine learning have shown interest in prediction markets, their focus thus far seems to concentrate on properties of market-maker mechanisms (e.g. Chen, Y. & Vaughan, 2010).

"pass", based on information about the game situation just before the play occurred. We chose the domain of making predictions about football games, in part, because it was analogous to more complex real-world predictions such as what actions would a business competitor take next. It also enabled us to emulate a realistic situation where humans and agents would have access to different information (specifically, humans had access to video information that is difficult or costly to codify for the agents).We hypothesized that 'hybrid' markets of humans and computers would do better than markets of either only computer-agents or only humans.

We conducted 20 laboratory sessions in which groups of 15 – 19 human subjects participated in prediction markets, both with and without computer agents. Overall there were 351 subjects, recruited from the general public via web advertising. For running prediction markets, we used the Zocalo open source software platform (available at http://zocalo.sourceforge.net/) with its logarithmic market maker, and with a custom version of the GUI that simplified trading for people less familiar with stock markets.

We used standard 3-layer artificial neural-net agents, developed using the JOONE open-source package (available at http://sourceforge.net/projects/joone/). For each play, the agents had three pieces of previously coded information: the down number, the number of yards to first down, and whether the previous play was a run or pass. We used a sigmoid transform function for our output layer, which limits the output to within the range of 0 to 1.

The agents were trained on a similar dataset of plays from a previous game. During the tuning phase, one agent was designed to make a trade in every market, no matter its confidence. It traded in the direction of the neural network output (i.e. output >= 0.5 means pass and output < 0.5 means run). Then an additional parameter, *BiasForAction*, was added to control the agent trading, such that agents traded only when their confidence level was above a threshold. A sensitivity analysis of the agents' performance with different values of bias for action (ranging from 0.01 to 0.3) allowed us to achieve correct classification for 87.5% of the plays in the test set when *BiasForAction* was set at 0.08. However, in that case the agents would trade in only about 25% of the markets. In order to have more agent participation, therefore, we set *BiasForAction* to 0.3 for the experiments reported here. This allowed the agents to participate in all markets, with 75% of the plays in the test set correctly classified.

There are, of course, many other possible kinds of artificial intelligence approaches that could be used here, many of which could presumably make the agents more accurate. As noted above, however, our goal was not to create the best possible artificial intelligence agents, merely to create one example of such agents for experimentation. What would have happened had we used better agents? Of course we don't know for sure, but it seems quite plausible that the overall pattern of results would remain the same. It is certainly possible that we would have gotten more accurate overall predictions. But regardless of whether humans, or agents, are better, we would expect the hybrid market to somewhat efficiently aggregate the knowledge of both. If more accurate agents enter the market, we would expect that human forecasters would participate less in the market, as they will have fewer opportunities to earn. And vice versa. In other words, we would expect people to intervene more when they believe the market (including other humans and agents) is wrong, and less, when they don't.

After initial explanation and training rounds, each experimental session included 20 plays from the same football game. For each play, a short video excerpt from the game was shown to participants. The video was automatically stopped just before the play was about to start. Then, an online prediction market was opened, and the participants (either humans only, or humans plus AI agents) started trading contracts of RUN and PASS (other plays were eliminated from the video). After 3.5 minutes of trading, the market was closed, and the video continued. The video revealed what play had actually occurred and then continued until just before the next play. The balance in each player's account was constantly updated with every trade and with every time the market closed, and carried over from one round to another. Compensation to participants included a base payment and an additional performance-based bonus that was proportional to the ending balance in each participant's account at the end of the experiment, and could reach up to 75% of the base pay.

In addition we ran 10 "computer-only" experimental sessions with 10 computer agents each. In these sessions, the agents traded with each other in separate markets for each of the 20 plays. We thus got a total of 600 observations: 10 observations for 20 plays in 3 conditions (humans only, computers only, and hybrid of humans and computers).

Results

As we predicted, the Hybrid markets were the most accurate, followed by Agents-only and then Humans-only (see **Table 1**). These differences were statistically significant ($p<0.05$) for the Log Scoring Rule (LSR). Statistical significance was not tested for the Mean Square Error (MSE) scoring rule because these scores were not normally distributed.

Measures of accuracy alone do not provide sufficient information to convey the complexity of the data; it is important to consider both the accuracy and the variability

of the errors. To do this, we used the ex post Sharpe ratio (Sharpe, 1994), originally developed to compare reward-to-risk performance of financial investments. To keep with the familiar logic of the Sharpe ratio, where higher positive returns are better, we adjust our scoring rules such that the adjusted MSE score (AMSE) equals 1-MSE. The adjusted Log score is $\log_{10}(P)$ where P is the prediction (market closing price) of the actual outcome. As a simple and straightforward benchmark, we use an "ignorant" predictor who bets 50% PASS all the time (and whose error variance is therefore zero). The corresponding AMSE and ALSR for the benchmark predictor are therefore 0.75 and 1.70, respectively. We summarize the results in **Table 1**. Clearly, the hybrid markets yield the highest Sharpe ratio, which means they offer a better tradeoff between prediction accuracy and error variability.

Table 1 - Accuracy and ex post Sharpe ratio results by type of prediction markets

	Accuracy		Sharpe Ratio	
Market Participants	MSE	LSR	AMSE*	ALSR**
Humans only	0.19	0.25	0.41	0.41
Agents only	0.17	0.23	0.39	0.37
Hybrid	0.15	0.21	0.74	0.72

*(Benchmark: 0.75) **(Benchmark: 1.70)

Our comparisons of accuracy, and of the Sharpe ratio, both rely on attaching values to prediction errors using scoring rules. While common, these rules may not represent the actual economic value of predictions (or the actual cost of corresponding errors), and in reality, it is not always possible to determine those values. The Receiver-Operating-Characteristic (ROC) is an established methodology for evaluating and comparing the performance of diagnostic and prediction systems (Swets, 1988), which does not rely on their unknown economic value, and hence, can provide additional support for our conclusions. ROC curves are obtained by plotting the hit rate (i.e., correctly identified events) versus the false alarm rate (incorrect event predictions) over a range of different thresholds that are used to convert probabilistic forecasts of binary events into deterministic binary forecasts. The area under the curve serves as a measure of the quality of the predictions, with a perfect predictor scoring 1. The ROC curves of our conditions are presented in **Figure 1** and the areas under the curves are depicted in **Table 2**. This result echoes our other findings, and yet again, suggests that the hybrid markets were more robust.

Figure 1: ROC Plots for Study 1

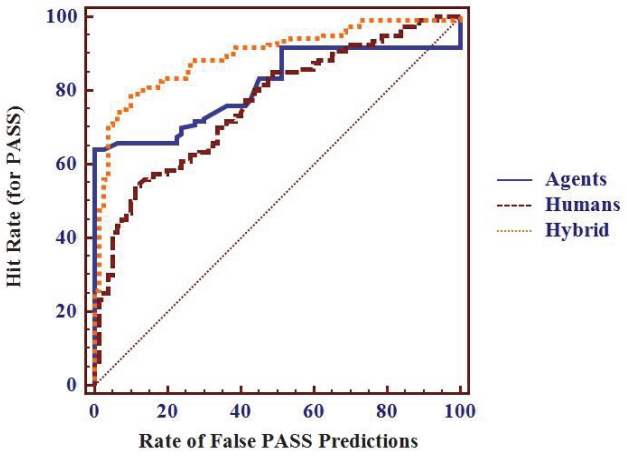

Table 2: Area under the ROC curves – all three conditions

	Area under ROC Curve	SE[3]
Humans-only Markets	0.76	0.03
Agents-only Markets	0.81	0.03
Hybrid Markets	0.90	0.02

Discussion

Predicting plays in a football game is an instance of a class of situations where the rules that determine a team's choice of action are fuzzy or difficult to discern, and where some data about the context are hard to codify. In this context, while on average the agents were more accurate than humans, they had a higher number of big errors. This may have been due, in part, to the fact that the agents did not have access to as much information about the game as the humans did. For instance, informal interviews with the human subjects revealed that they indeed used information from the video that would have been difficult to code for agents (such as team formation, body language, and announcers' comments).

The combination of humans and agents provided predictions that were more calibrated than those of the agents, more discriminating than those of the humans, and overall providing a better tradeoff of calibration and discrimination compared to the humans or the agents. Predictions made by hybrid humans-and-agents markets also provided the best tradeoff of accuracy and variability of prediction errors, as measured by the Sharpe ratio. An

[3] Standard errors were calculated using the method offered by DeLong, et al. (DeLong, DeLong, & Clarke-Pearson, 1988). However they may be inaccurate as we used repeated measurements.

ROC analysis, which does not rely on any assumptions about the cost of errors, also shows that hybrid markets provide a better trade-off between good and bad predictions. Overall, therefore, the combination of human and agent predictions proved more robust, and arguably, superior to either the agents-only predictions or the humans-only predictions in our setting.

We thus provide a proof of concept of the existence of scenarios where combining predictions from groups of humans and artificial-intelligence agents can outperform groups of either type alone. We also show that prediction markets provide a useful way to combine predictions from humans and models, providing what we believe to be the first systematically studied attempt at using them for this purpose. As discussed above, many different ways for combining predictions (from either humans or models) are explored in the literature, and no single method is best (or even suitable) for all settings. It is not our goal to argue that prediction markets are always the best method for combining multiple predictions, but they are appealing for a number of reasons described in the Introduction. Our study further shows that prediction markets can produce valuable predictions in a very short time (minutes), thus potentially serving as a real-time prediction tool. In this sense, our work responds to calls in the literature on predictive analytics and business intelligence for reactive components to help decision makers monitor and respond to time-critical operational processes (e.g. Matteo, Stefano, & Iuris, 2004).

As previous research has shown, there are many business and other situations where mechanical predictions based on structured data are superior to predictions made by humans (Grove, et al., 2000). On the other hand, there are many other situations where the factors relevant to predicting are so complex, or where there is so little codifiable data, that no suitable mechanical models even exist; in which case, the only option is to rely on human judgment.

But we believe there are also many important real-world situations where combining predictions from humans and agents can be valuable. For instance, this approach may be particularly beneficial in complex situations involving the actions of human groups (such as business competitors, customers, partners, or regulators), where it may be difficult to discern or formulate all the rules governing the phenomena of interest but where there is still a fair amount of relevant data that can be analyzed. One prototypical example of such a situation, for instance, would be predicting sales of fashion-driven products. There is a great deal of relevant data that can be analyzed, but some important factors are very difficult to quantify in advance. In such domains, machine learning and other quantitative methods can be useful in building sophisticated and adaptive models based on potentially vast amounts of data (for recent examples, see Bohorquez, Gourley, Dixon, Spagat, & Johnson, 2009; Mannes, et al., 2008); and humans' tacit knowledge, ability to acquire unstructured information, and intuition can help in both information retrieval, and preventing catastrophic prediction errors.

We believe that the work reported here shows the potential value of combining predictions from humans and agents in situations with complex rules and difficult-to-codify information. Additional work is required to identify and compare other ways of combining human and machine predictions, and to understand their respective advantages and disadvantages in different contexts. Future work should also examine this approach in more complex domains, and with more sophisticated, domain-specific agents. We hope our initial work will encourage others to further investigate this promising direction.

Acknowledgments

We thank MIT Lincoln Laboratory and the U.S. Army Research Laboratory's Army Research Office (ARO) for funding this project. We are grateful to John Willett for his patience, dedication and help with statistical analyses and to Chris Hibbert for software development, and education about prediction markets. We thank Sandy Pentland, Tomaso Poggio, Drazen Prelec, and Josh Tenenbaum for many discussions out of which this project originated, and benefited greatly. For help with software and experimental design we thank Jason Carver, Wendy Chang, Jeremy Lai and Rebecca Weiss. For their wise comments we thank John Carroll, Gary Condon, Robin Hanson, Haym Hirsh, Josh Introne, Ben Landon, Retsef Levi, David Pennock, Cynthia Rudin, and Paulina Varshavskaya. This paper also benefited from constructive comments of participants and two anonymous reviewers of the NIPS 2010 Crowdsourcing and Computational Social Science Workshop, as well as those of the associate editor and two anonymous reviewers of ICIS 2011. Thanks also go to our research assistants Jonathan Chapman, Catherine Huang, Natasha Nath, Carry Ritter, Kenzan Tanabe and Roger Wong, and to Richard Hill and Robin Pringle for administrative support.

References

Armstrong, J. S. (1989). Combining forecasts: The end of the beginning or the beginning of the end? *International Journal of Forecasting, 5*, 585-588.

Armstrong, J. S. (2001a). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researchers and practitioners*: Kluwer Academic Publishers.

Armstrong, J. S. (2001b). JUDGMENTAL BOOTSTRAPPING: INFERRING EXPERTS' RULES FOR FORECASTING. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic Publishers.

Berg, J. E., Forsythe, R., Nelson, F., & Rietz, T. A. (2001). Results from a Dozen Years of Election Futures Markets Research. *Handbook of Experimental Economic Results*, 486–515.

Berg, J. E., Nelson, F. D., & Rietz, T. A. (2008). Prediction market accuracy in the long run. *International Journal of Forecasting, 24*(2), 283-298. doi: 10.1016/j.ijforecast.2008.03.007

Blattberg, R. C., & Hoch, S. J. (1990). Database Models and Managerial Intuition: 50% Model+ 50% Manager. *Management Science, 36*(8), 887-899.

Bohorquez, J. C., Gourley, S., Dixon, A. R., Spagat, M., & Johnson, N. F. (2009). Common ecology quantifies human insurgency. *Nature, 462*(7275), 911-914.

Brown, R. (1986). *Social psychology* (2nd ed.). New York, NY: Free Press.

Bunn, D., & Wright, G. (1991). Interaction of judgemental and statistical forecasting methods: Issues and analysis. *Management Science, 37*(5), 501-518.

Chen, K.-Y., & Plott, C. R. (2002). Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem. *California Institute of Technology, Division of the Humanities and Social Sciences, Working Paper 1131*.

Chen, Y., Chu, C.-H., Mullen, T., & Pennock, D., M. . (2005). *Information markets vs. opinion pools: an empirical comparison*. Paper presented at the Proceedings of the 6th ACM conference on Electronic commerce, Vancouver, BC, Canada.

Chen, Y., & Vaughan, J. W. (2010). *A new understanding of prediction markets via no-regret learning.* Proceedings of the 11th ACM conference on Electronic Commerce (EC '10), Cambridge, MA (pp. 189-198.)

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559-583.

Cowgill, B., Wolfers, J., & Zitzewitz, E. (2008). Using Prediction Markets to Track Information Flows: Evidence from Google. *Dartmouth College*.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics, 44*(3), 837-845.

Duin, R., & Tax, D. (2000). *Experiments with classifier combining rules.* Proceedings of the First International Workshop on Multiple Classifier Systems (MCS 2000), Cagliari, Italy (pp. 16-29.)

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7*(1), 86-106.

Goel, S., Reeves, D. M., Watts, D. J., & Pennock, D. M. (2010). *Prediction Without Markets.* Proceedings of the 11th ACM conference on Electronic commerce, Cambridge, MA (pp. 357-366.)

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73*(6), 422-432.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30.

Hayek, F. A. (1945). The Use of Knowledge in Society. *The American Economic Review, 35*(4), 519-530.

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting, 21*(1), 15-24.

Hopman, J. (2007). Using forecasting markets to manage demand risk. *Intel Technology Journal, 11*, 127–136.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 4-37.

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*: Houghton Mifflin Boston.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review, 80*(4), 237-251.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: toward an integrative approach. *Psychological Bulletin, 107*(3), 296-310.

Kuncheva, L. I. (2002). A Theoretical Study on Six Classifier Fusion Strategies. *IEEE Transactions on pattern analysis and machine intelligence, 24*(2), 281-286. doi: 10.1109/34.982906

Lam, L. (2000). *Classifier combinations: implementations and theoretical issues.* Proceedings of the First International Workshop on Multiple Classifier Systems (MCS 2000), Cagliari, Italy (pp. 77-86.)

Lichtenstein, S., Baruch, F., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*: Cambridge University Press.

Makridakis, S. (1989). Why combining works? *International Journal of Forecasting, 5*(4), 601-603.

Malone, T. W. (2004). Bringing the market inside. *Harvard Business Review, 82*(4), 106-114.

Mannes, A., Michael, M., Pate, A., Sliva, A., Subrahmanian, V. S., & Wilkenfeld, J. (2008). Stochastic Opponent Modeling Agents: A Case Study with Hezbollah. In H. Liu, J. J. Salerno & M. J. Young (Eds.), *Social Computing, Behavioral Modeling, and Prediction* (pp. 37-45).

Matteo, G., Stefano, R., & Iuris, C. (2004). *Beyond data warehousing: what's next in business intelligence?* Paper presented at the Proceedings of the 7th ACM international workshop on Data warehousing and OLAP, Washington, DC, USA.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Mitra, N., J., Chu, H.-K., Lee, T.-Y., Wolf, L., Yeshurun, H., & Cohen-Or, D. (2009). *Emerging images*. Paper presented at the ACM SIGGRAPH Asia, Yokohama, Japan.

Ortner, G. (1997). *Forecasting Markets—An Industrial Application*. University of Technology Vienna.

Perols, J., Chari, K., & Agrawal, M. (2009). Information market-based decision fusion. *Management Science, 55*(5), 827-842.

Sharpe, W. F. (1994). The Sharpe ratio. *Journal of portfolio management* (Fall), 49-58.

Spann, M., & Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting, 28*(1), 55-72.

Stewart, T. R. (2001). Improving reliability of judgmental forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 81-106): Kluwer Academic Publishers.

Sunstein, C. R. (2006). *Infotopia: How Many Minds Produce Knowledge*: Oxford University Press, USA.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.

Tziralis, G., & Tatsiopoulos, I. (2007). Prediction Markets: An Extended Literature Review. *Journal of Prediction Markets, 1*(1), 75-91.

von Ahn, L., Blum, M., Hopper, N., & Langford, J. (2003). CAPTCHA: Using Hard AI Problems for Security. In *Advances in Cryptology — EUROCRYPT 2003* (Vol. 2656): Springer Berlin / Heidelberg.

von Ahn, L., & Dabbish, L. (2004). *Labeling images with a computer game*. Paper presented at the Proceedings of the ACM SIGCHI conference on Human factors in computing systems, Vienna, Austria.

Winkler, R. L. (1989). Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting, 5*(4), 605-609.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *The Journal of Economic Perspectives, 18*(2), 107-126.